



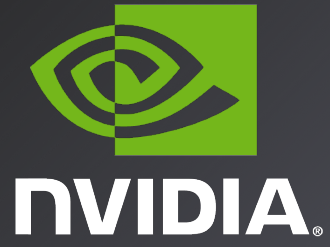
Flash Memory Summit

# Feeding Data Hungry GPUs with Networked Flash



**Chris Lamb**  
**VP Compute Software at NVIDIA**  
**Michael Kagan**  
**CTO at Mellanox**





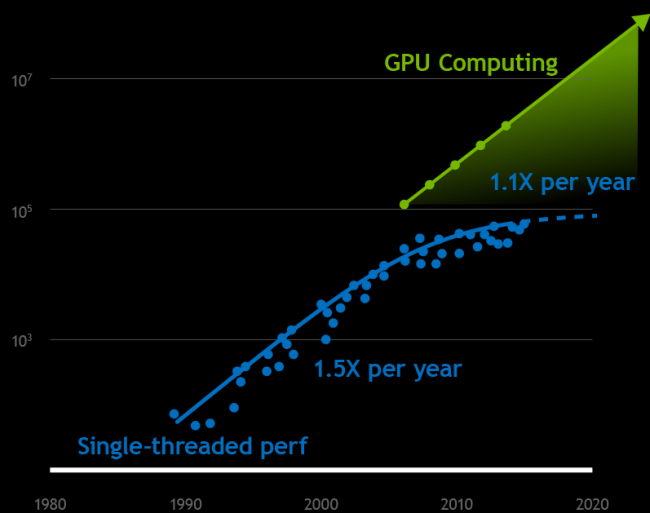
# FEEDING DATA HUNGRY GPUS WITH NETWORKED FLASH

Chris Lamb, VP Compute Software, NVIDIA

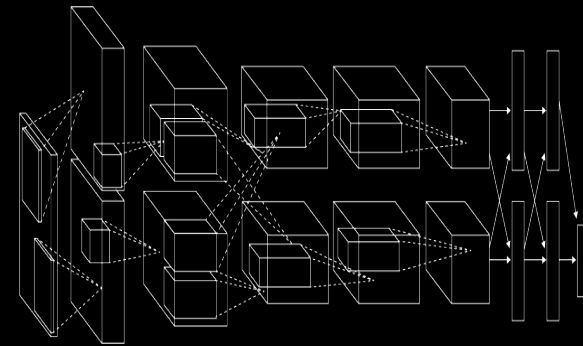


# TWO FORCES SHAPING COMPUTING

40 YEARS OF CPU TREND DATA



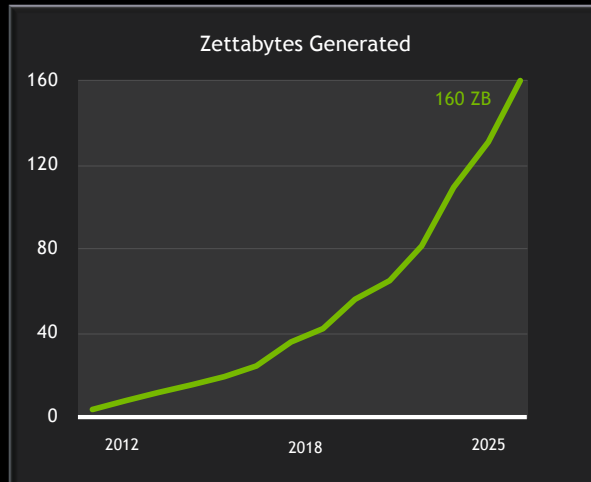
ALEXNET: THE SPARK OF THE MODERN AI ERA



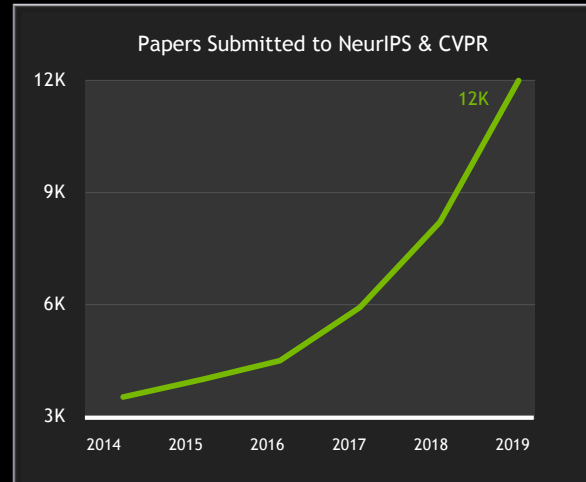
30-year era of Moore's law is ending  
Machines can learn with AI, but need massive computer power  
Optimizing GPU systems, software, and algorithms is the path to 1,000X speed-up

# EXPONENTIAL GROWTH IN COMPUTING DEMAND

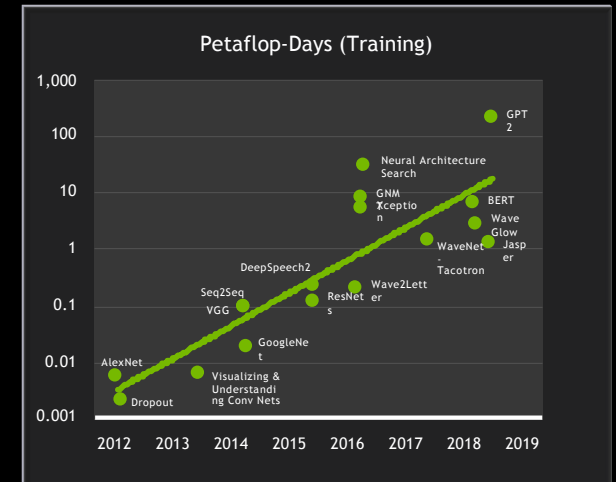
## DATA SIZE GROWING



## AI RESEARCH GROWING



## AI MODEL COMPLEXITY GROWING



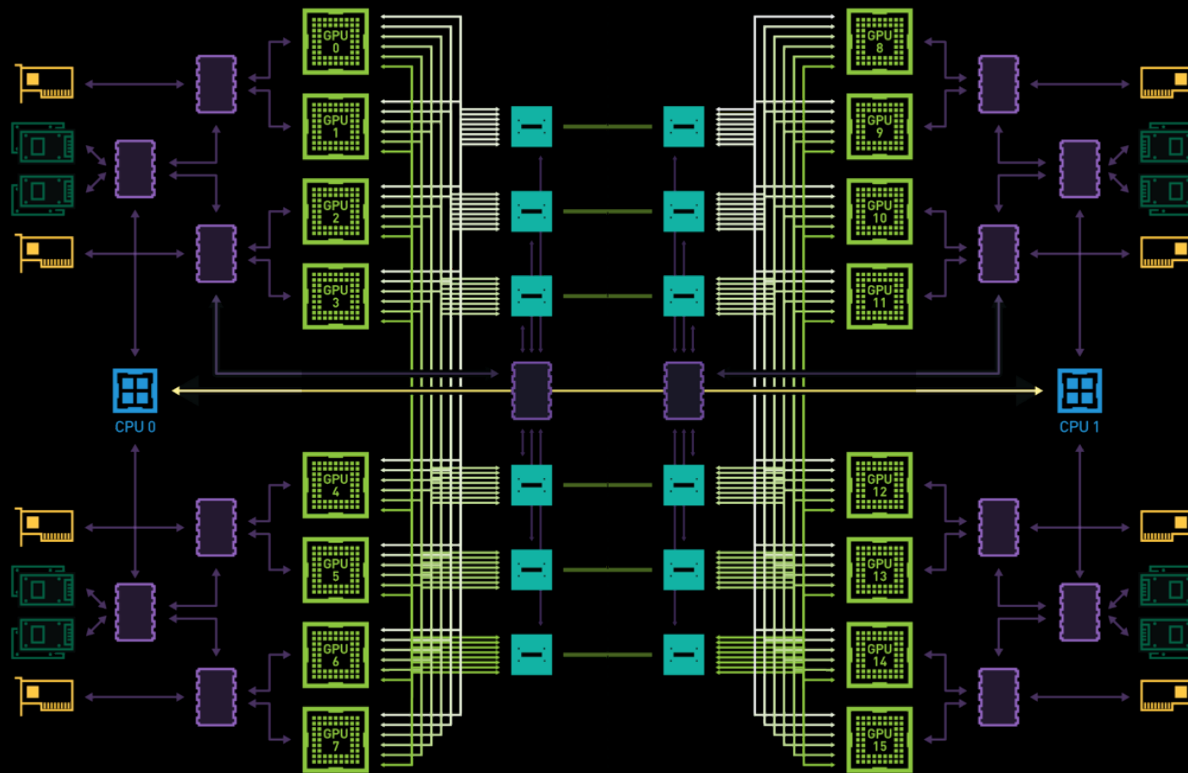
Source: IDC, GitHub, and OpenAI / NVIDIA

# NVIDIA CUDA-X IS THE ENGINE OF AI



CUDA-X AI Addresses the End-to-end Development of AI | Optimized For Every AI Model and Framework  
Available in Every Cloud and from Every Computer Maker

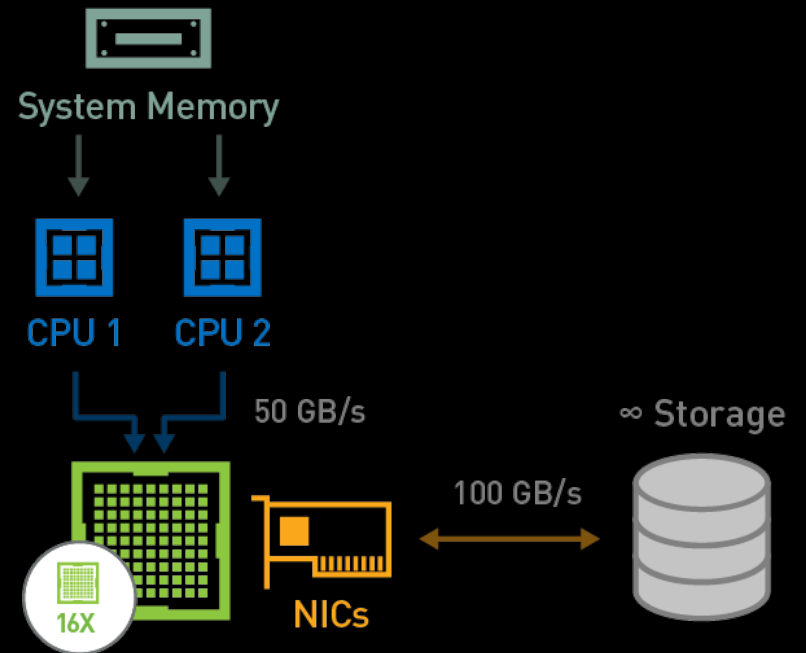
## 2.4 TB Bisection Bandwidth



Mellanox NICs NVMe PCIe Switches NVSwitch NVLink PCIe QPI

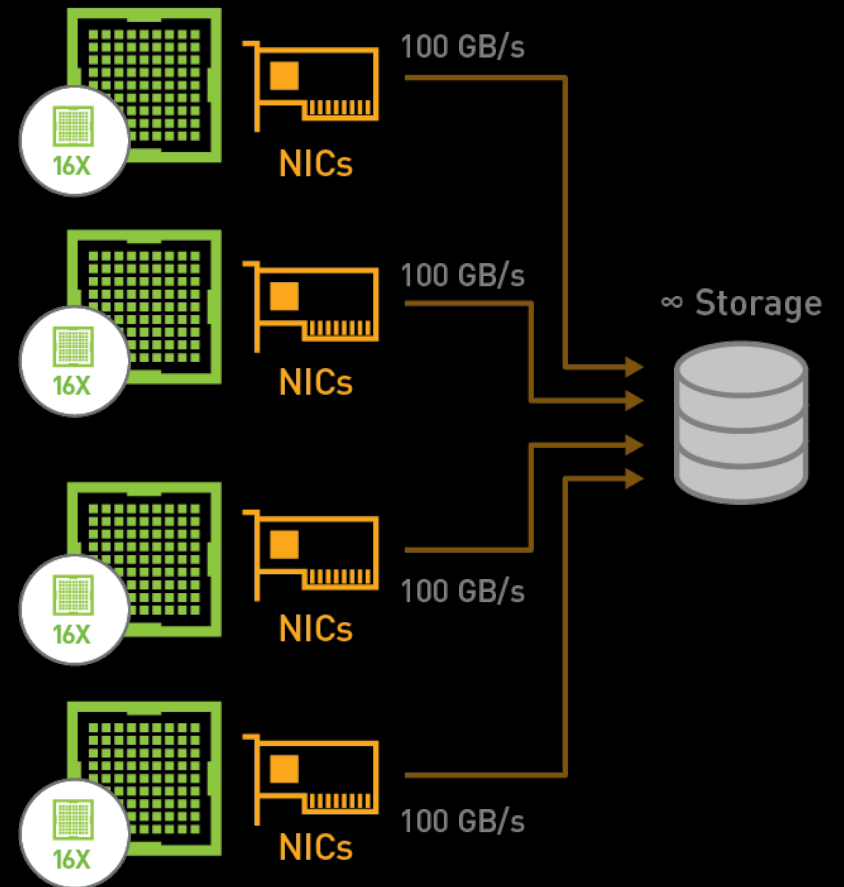
# GPU OPTIMIZED SYSTEMS

With GPUDirect Storage



# GPU OPTIMIZED DATA CENTERS

Clusters with GPUDirect Storage



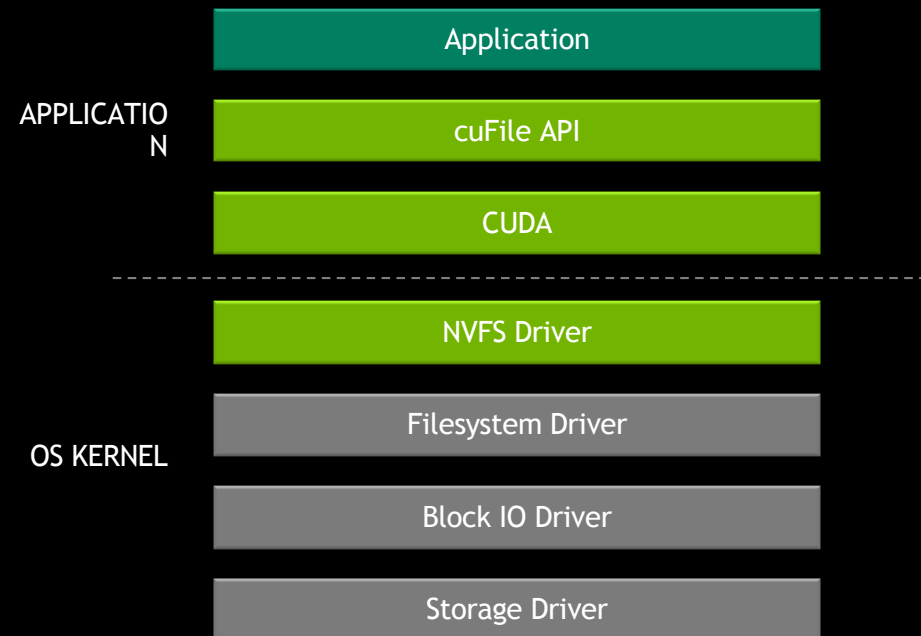


# CUFILE AND GPUDIRECT STORAGE

## Architecture of the Stack

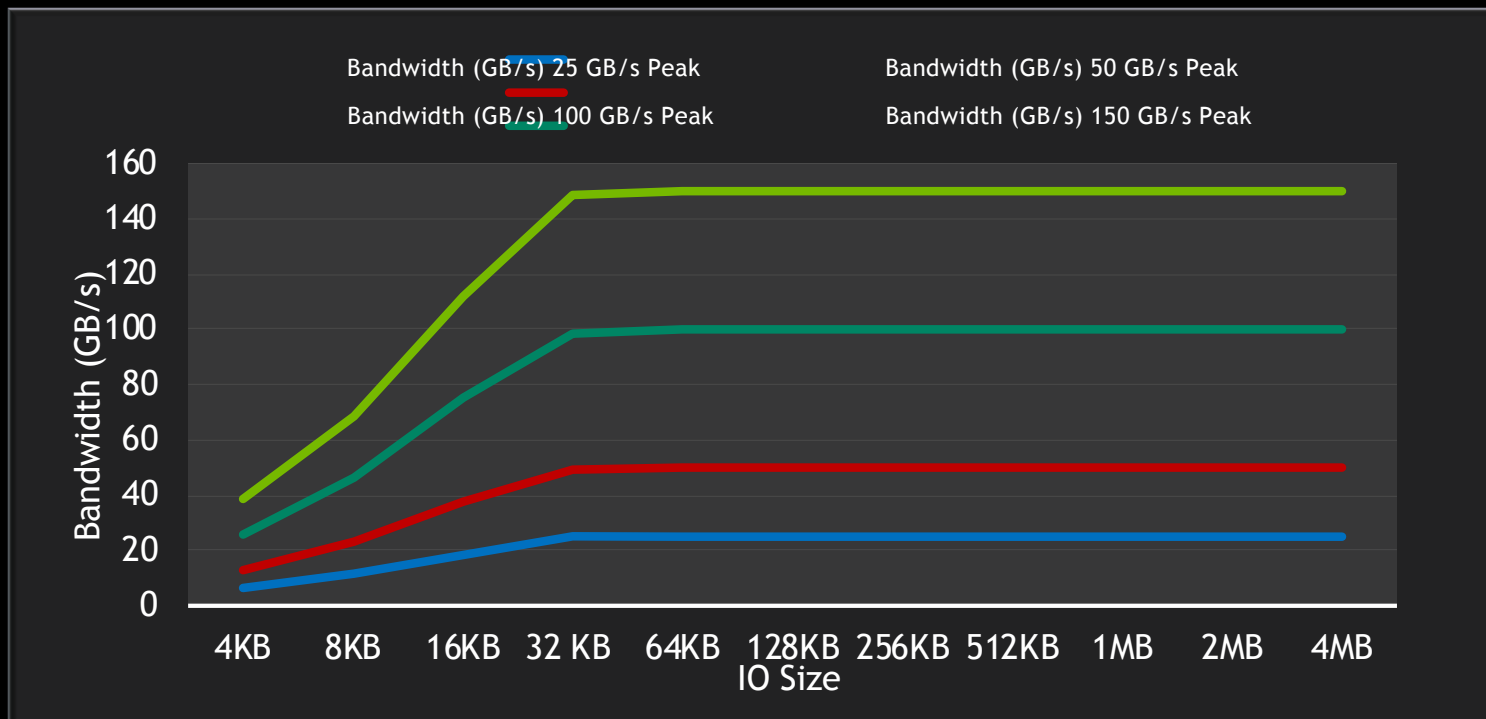
cuFile API  
For applications

NVFS Driver API  
For filesystem, block, and storage drivers



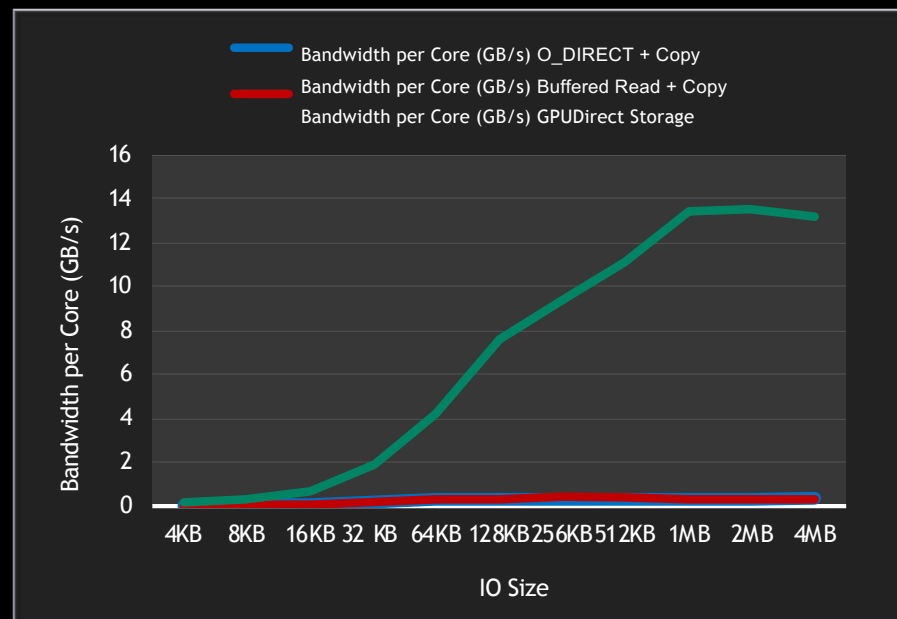
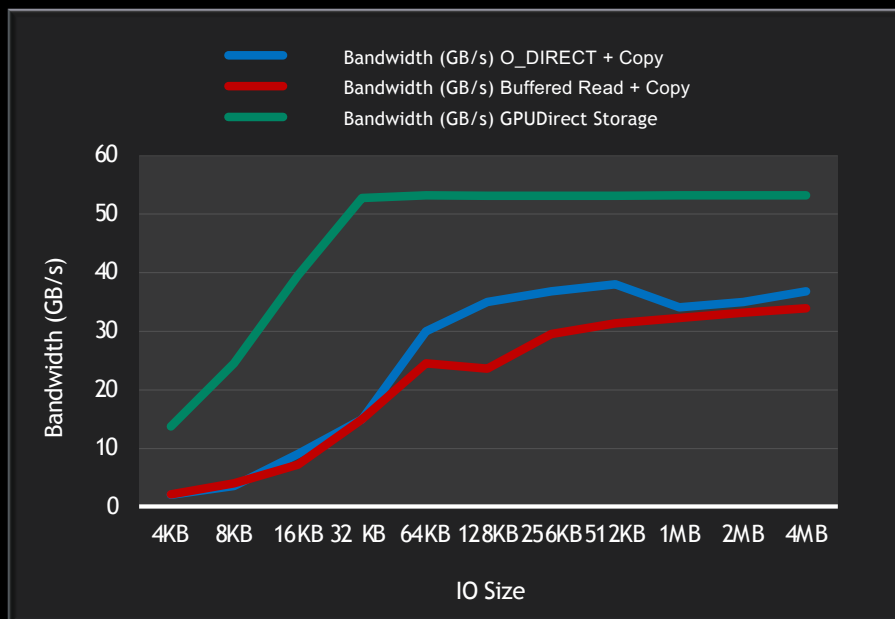
# BANDWIDTH SCALES TRANSPARENTLY

## Architectural Peak Bandwidth



# EXTREME FILE IO BANDWIDTH

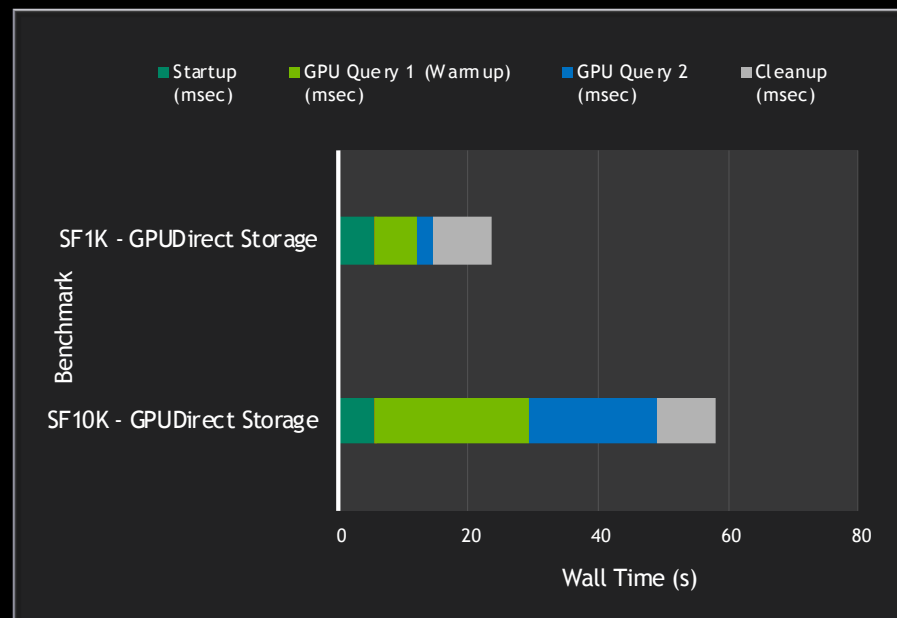
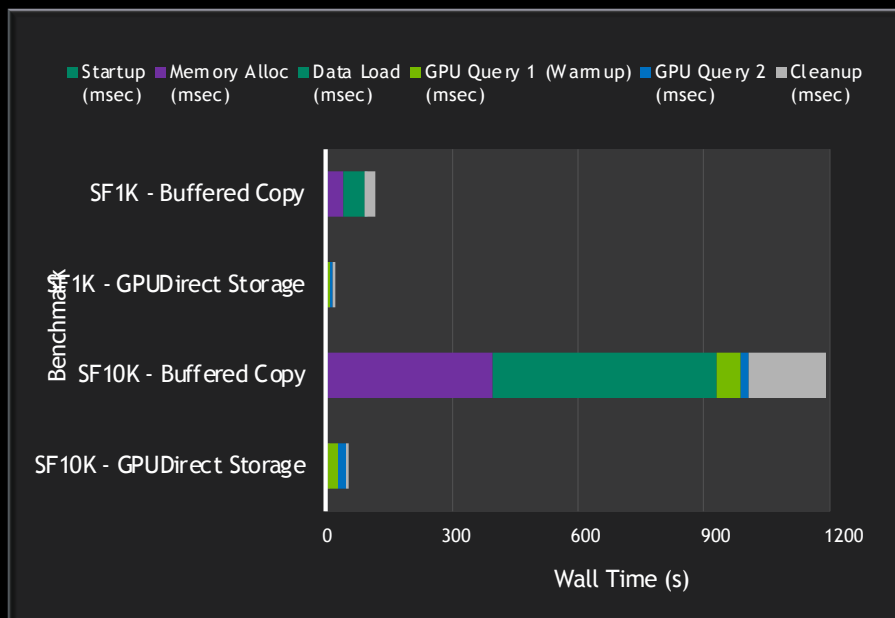
Up to 10x Bandwidth Per Core



FIO + GDS extension on EXT4 Filesystem - synchronous ordered IO with 64 job per device at queue depth 1

# SCALED APPLICATION PERFORMANCE

Up to 20x Faster with GPU Accelerated TPC-H Query 4 Scaling on EXT4



# FOR MORE INFORMATION

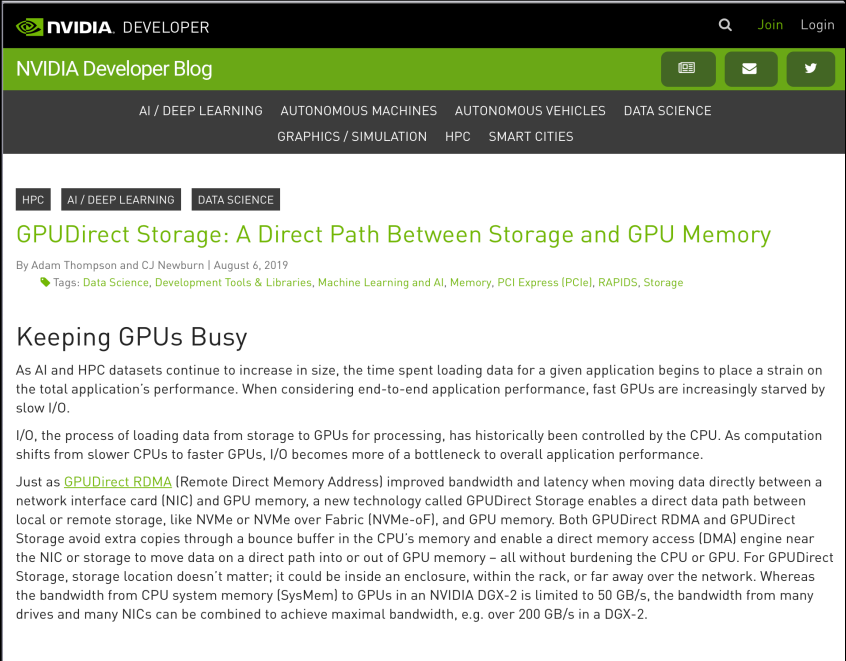
Join the GPUDirect Storage interest list in order to:

Provide feedback

Extend with other filesystems

Technical blog and link to sign up:

<https://devblogs.nvidia.com/gpudirect-storage/>



The screenshot shows the NVIDIA Developer Blog interface. At the top, there is a search bar and links for 'Join' and 'Login'. The main navigation bar includes categories like 'AI / DEEP LEARNING', 'AUTONOMOUS MACHINES', 'AUTONOMOUS VEHICLES', 'DATA SCIENCE', 'GRAPHICS / SIMULATION', 'HPC', and 'SMART CITIES'. The article title is 'GPUDirect Storage: A Direct Path Between Storage and GPU Memory' by Adam Thompson and CJ Newburn, dated August 6, 2019. The article is tagged with 'Data Science, Development Tools & Libraries, Machine Learning and AI, Memory, PCI Express (PCIe), RAPIDS, Storage'. The sub-header is 'Keeping GPUs Busy'. The main text discusses the performance impact of data loading on GPUs and introduces GPUDirect Storage as a solution to bypass the CPU bottleneck.

**NVIDIA DEVELOPER** Join Login

**NVIDIA Developer Blog** 📄 ✉ 🐦

AI / DEEP LEARNING AUTONOMOUS MACHINES AUTONOMOUS VEHICLES DATA SCIENCE  
GRAPHICS / SIMULATION HPC SMART CITIES

HPC AI / DEEP LEARNING DATA SCIENCE

## GPUDirect Storage: A Direct Path Between Storage and GPU Memory

By Adam Thompson and CJ Newburn | August 6, 2019

Tags: Data Science, Development Tools & Libraries, Machine Learning and AI, Memory, PCI Express (PCIe), RAPIDS, Storage

### Keeping GPUs Busy

As AI and HPC datasets continue to increase in size, the time spent loading data for a given application begins to place a strain on the total application's performance. When considering end-to-end application performance, fast GPUs are increasingly starved by slow I/O.

I/O, the process of loading data from storage to GPUs for processing, has historically been controlled by the CPU. As computation shifts from slower CPUs to faster GPUs, I/O becomes more of a bottleneck to overall application performance.

Just as [GPUDirect RDMA](#) (Remote Direct Memory Address) improved bandwidth and latency when moving data directly between a network interface card (NIC) and GPU memory, a new technology called GPUDirect Storage enables a direct data path between local or remote storage, like NVMe or NVMe over Fabric (NVMe-oF), and GPU memory. Both GPUDirect RDMA and GPUDirect Storage avoid extra copies through a bounce buffer in the CPU's memory and enable a direct memory access (DMA) engine near the NIC or storage to move data on a direct path into or out of GPU memory – all without burdening the CPU or GPU. For GPUDirect Storage, storage location doesn't matter; it could be inside an enclosure, within the rack, or far away over the network. Whereas the bandwidth from CPU system memory [SysMem] to GPUs in an NVIDIA DGX-2 is limited to 50 GB/s, the bandwidth from many drives and many NICs can be combined to achieve maximal bandwidth, e.g. over 200 GB/s in a DGX-2.





# Mellanox<sup>®</sup>

TECHNOLOGIES



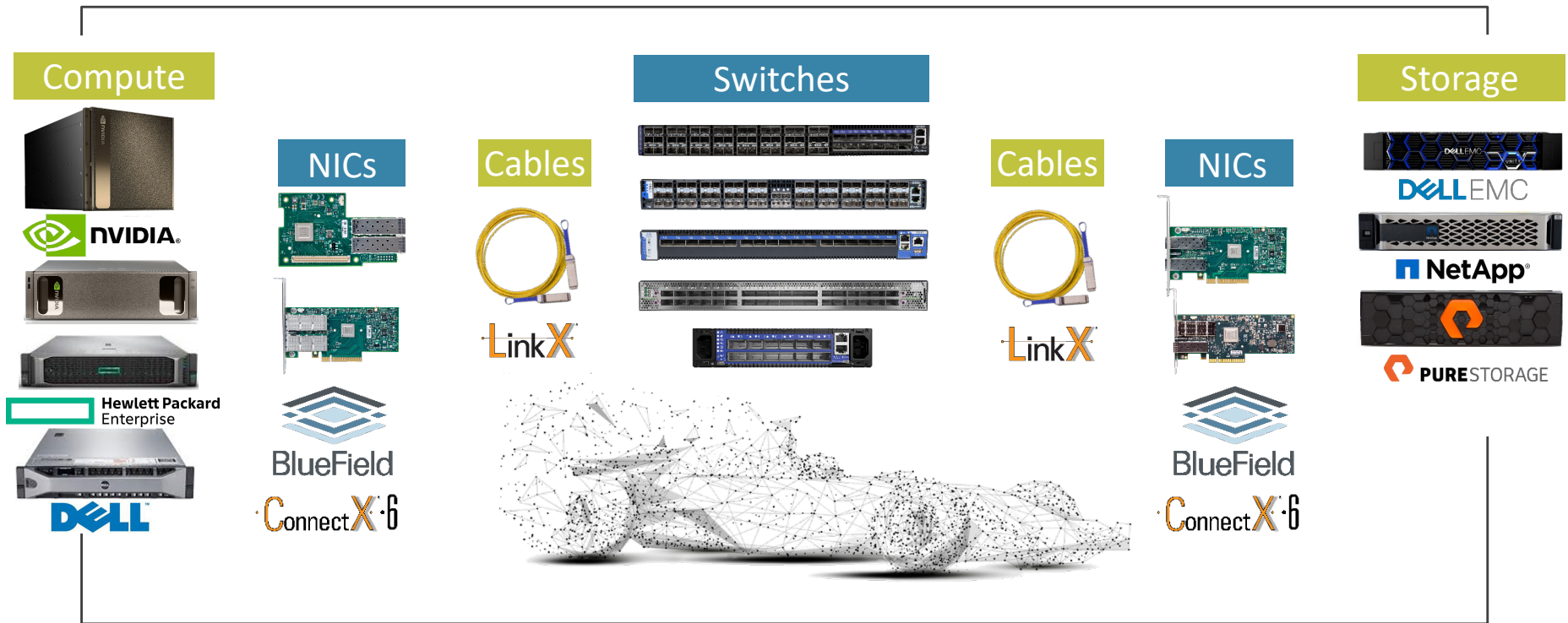
**Michael Kagan, CTO at Mellanox**





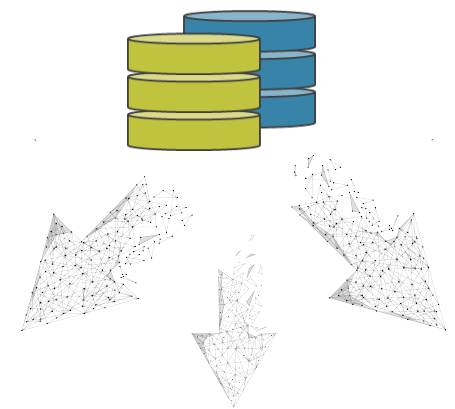
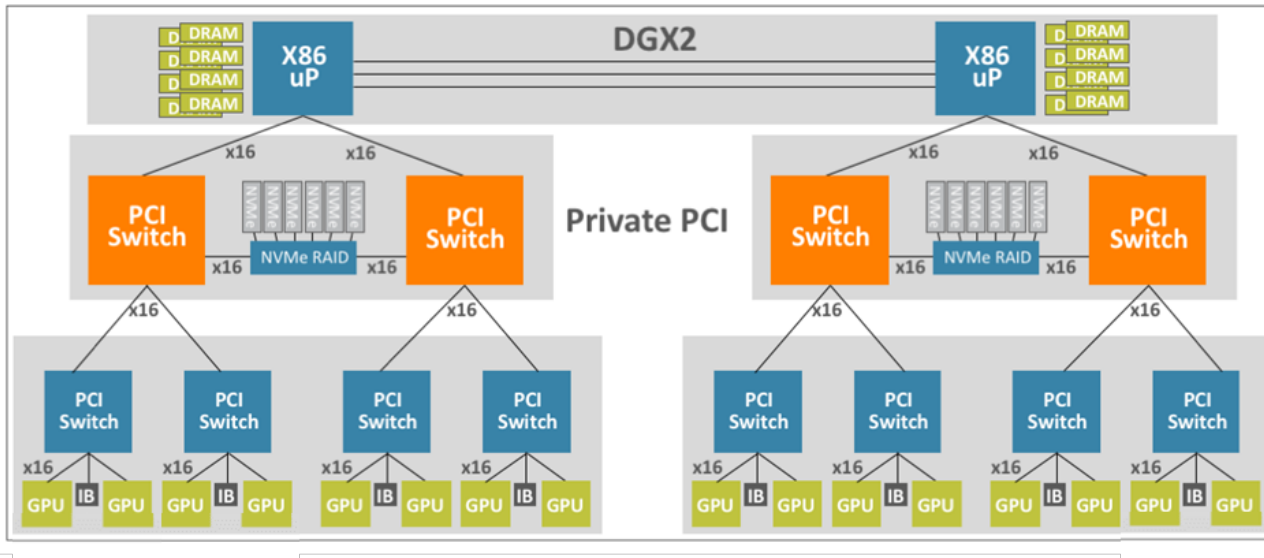
# Mellanox Networking Delivers the Solution

Industry Leading Ethernet & InfiniBand End-to-End  
25, 40, 50, 56, 100, 200Gb/s (and soon 400Gb/s)



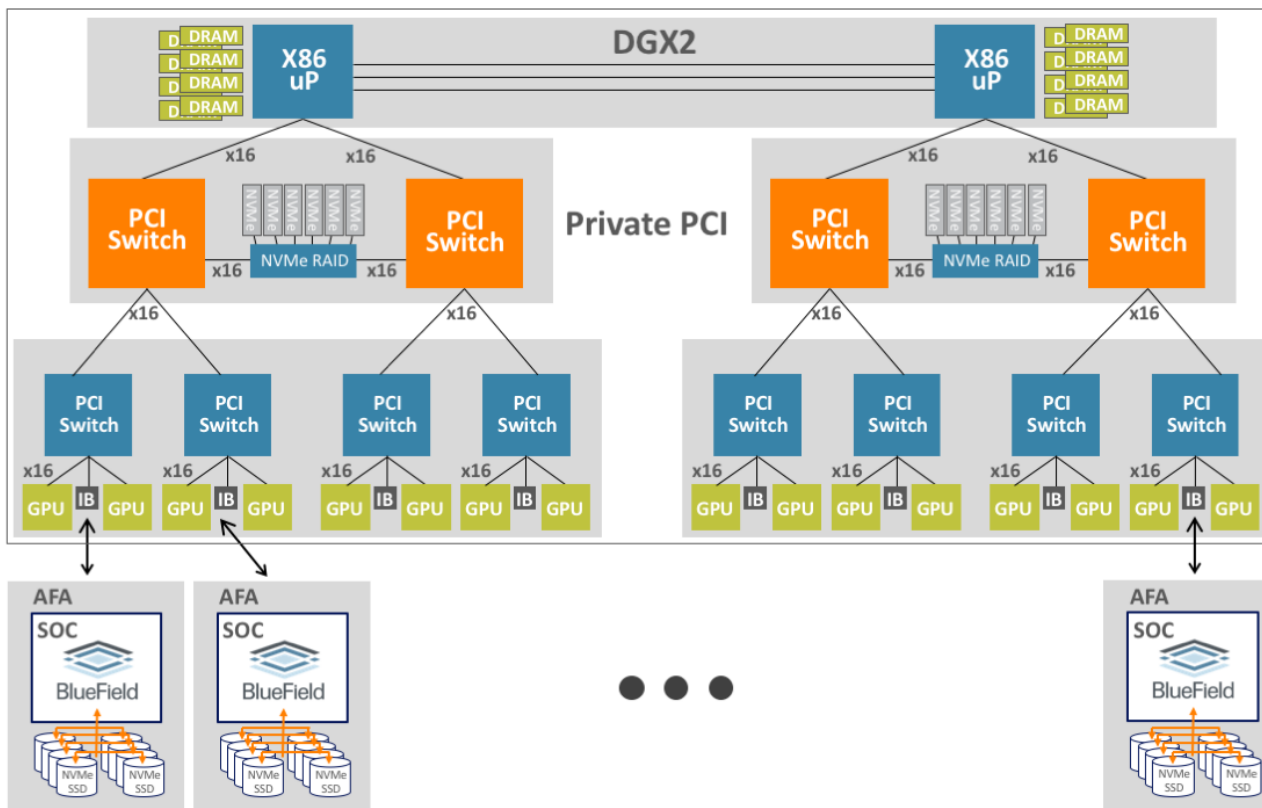


# 12.5GB/s of Storage to Every Tesla Pair

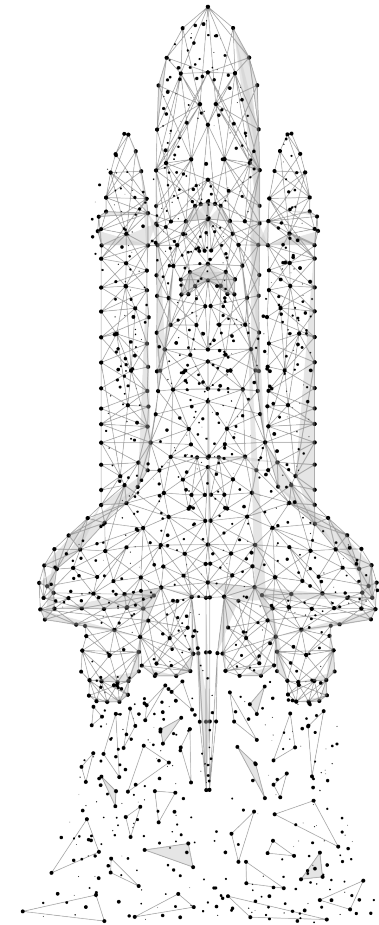


**Internal NVMe SSD storage**

# 25GB/s to Every Tesla Pair



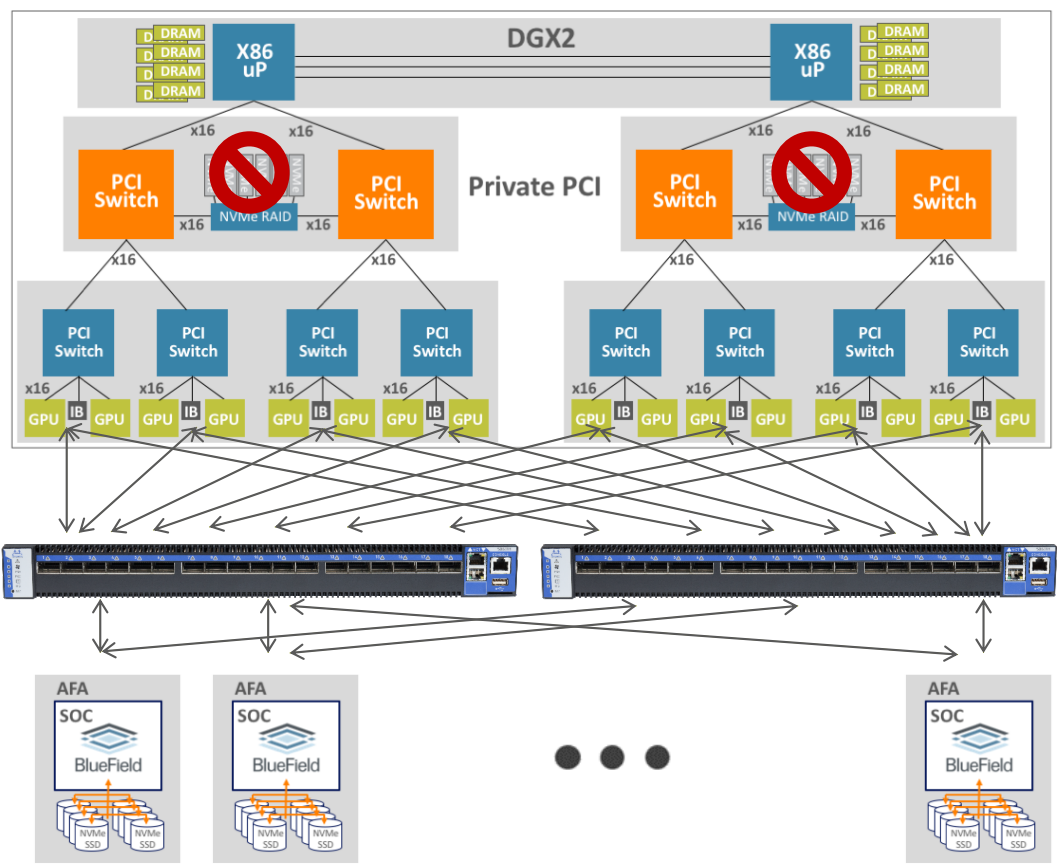
**External IB or Ethernet attached NVMe JBOFs**



# Unlimited High Performance Storage when Networked

## Plus: Storage Networking Advantages over Local Storage for GPUs

- Unlimited capacity
- High Availability
- Higher Utilization
- Lower TCO
  - Less Power
  - Less Rack Space
  - Easier Management
  - Less Cost



DGX

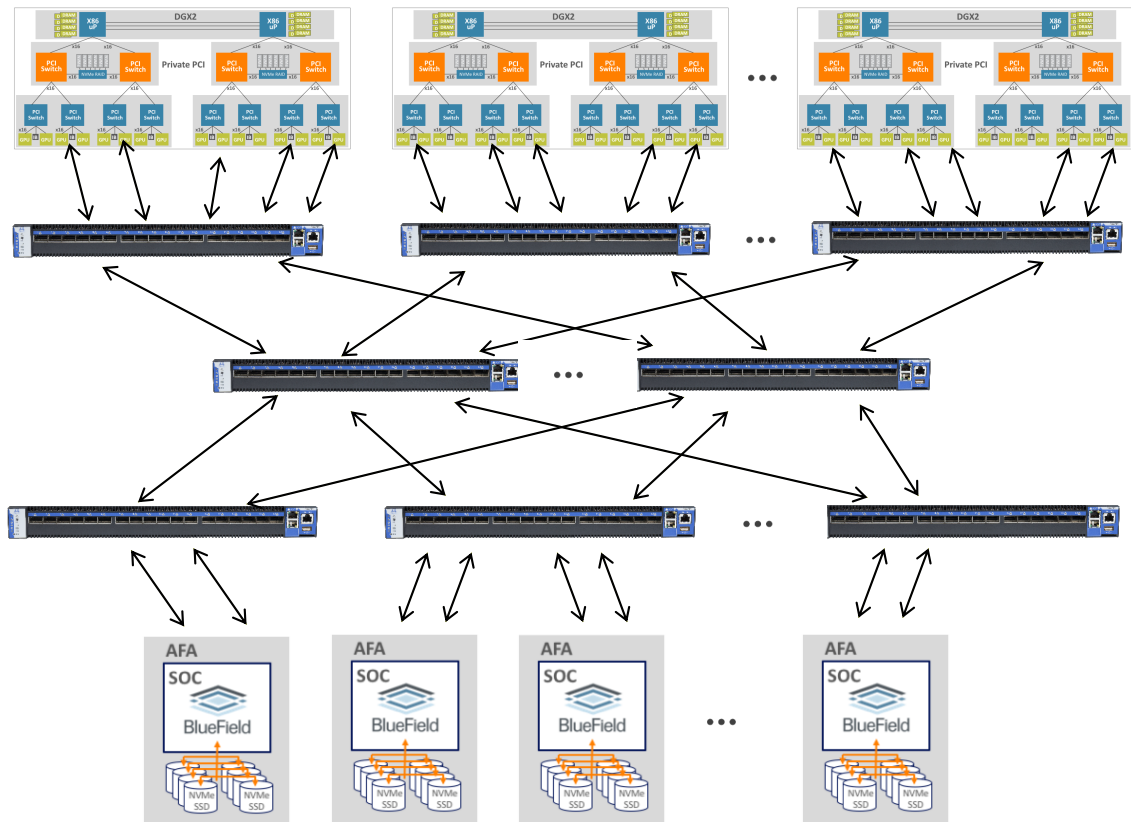
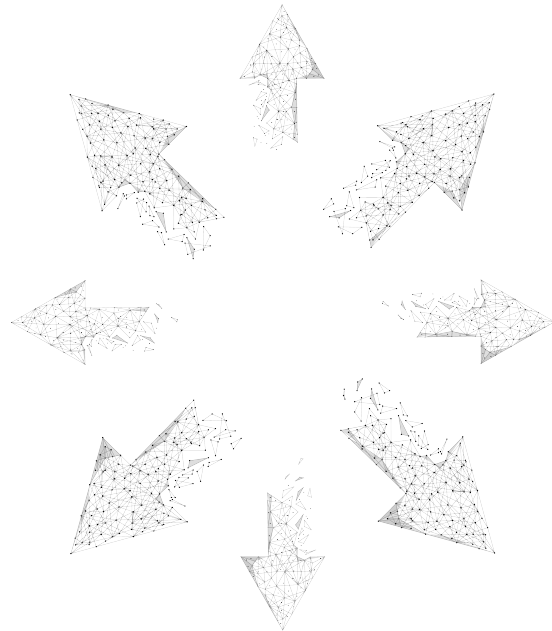
Dual port  
200Gb  
HCA or NICs

200Gb IB or  
Ethernet  
Switches

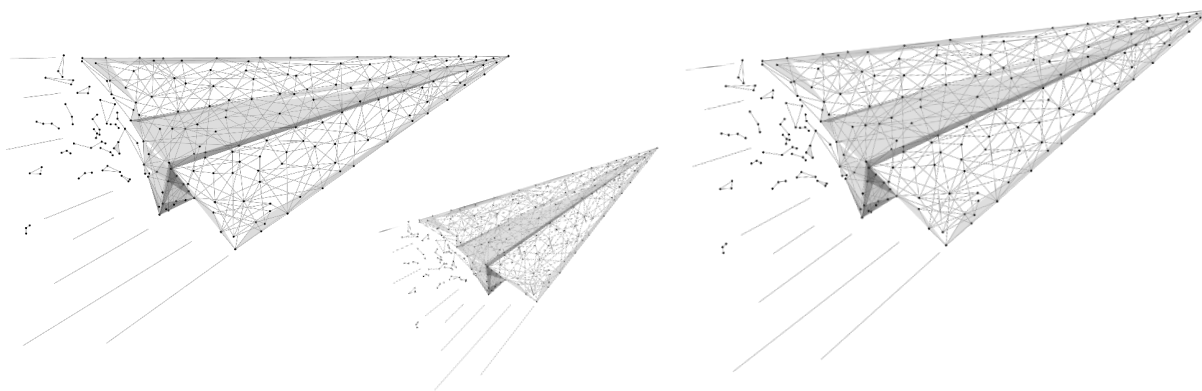
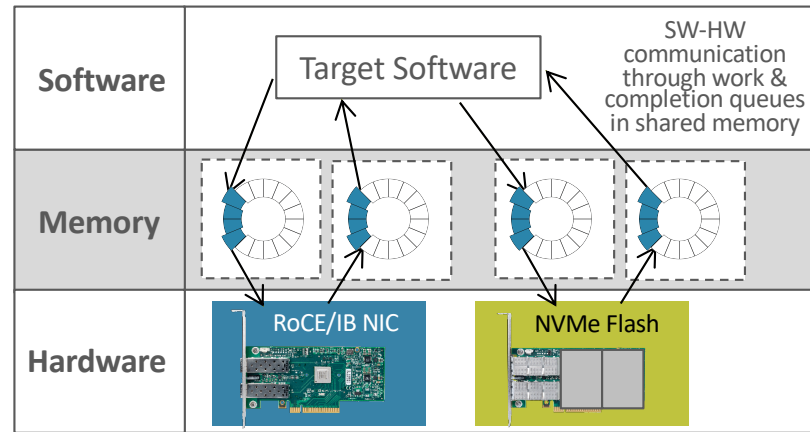
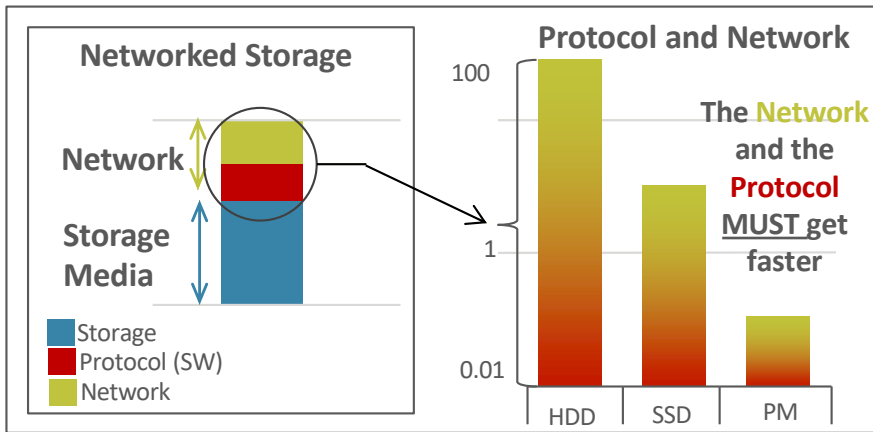
NVMe-of  
All Flash  
Arrays

# Now Storage & GPUs can Scale Independently

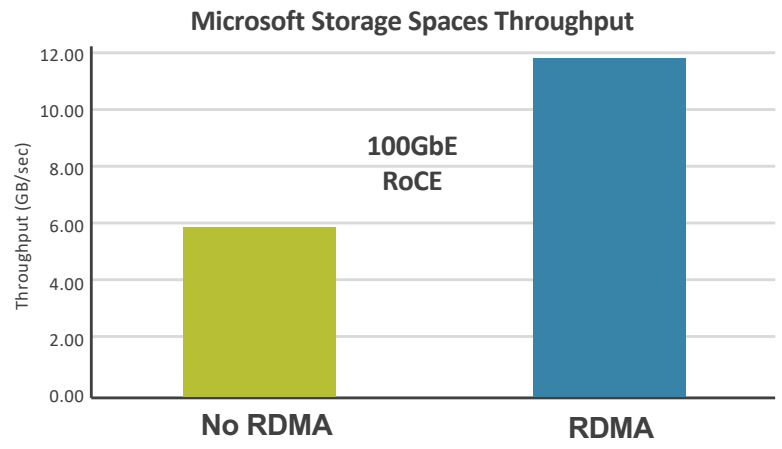
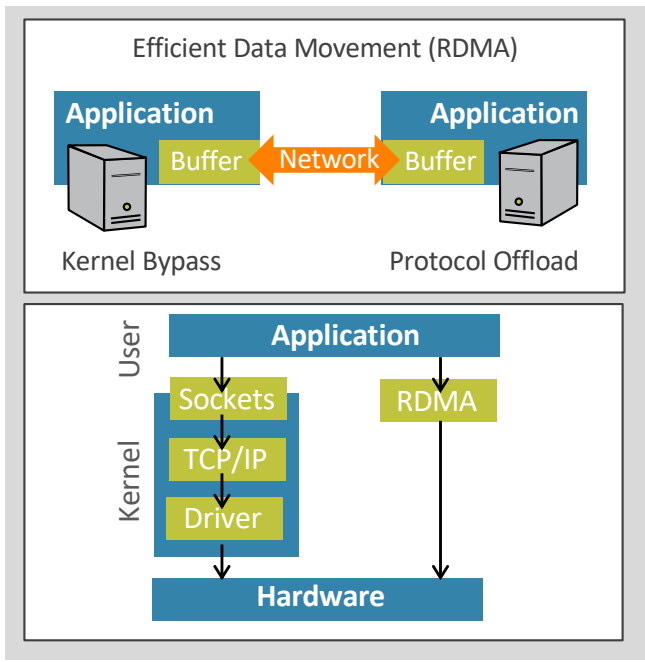
All because of Mellanox high performance ultra low latency storage networking



# NVMe, NVMe-oF & RDMA Protocols



# RDMA Performance



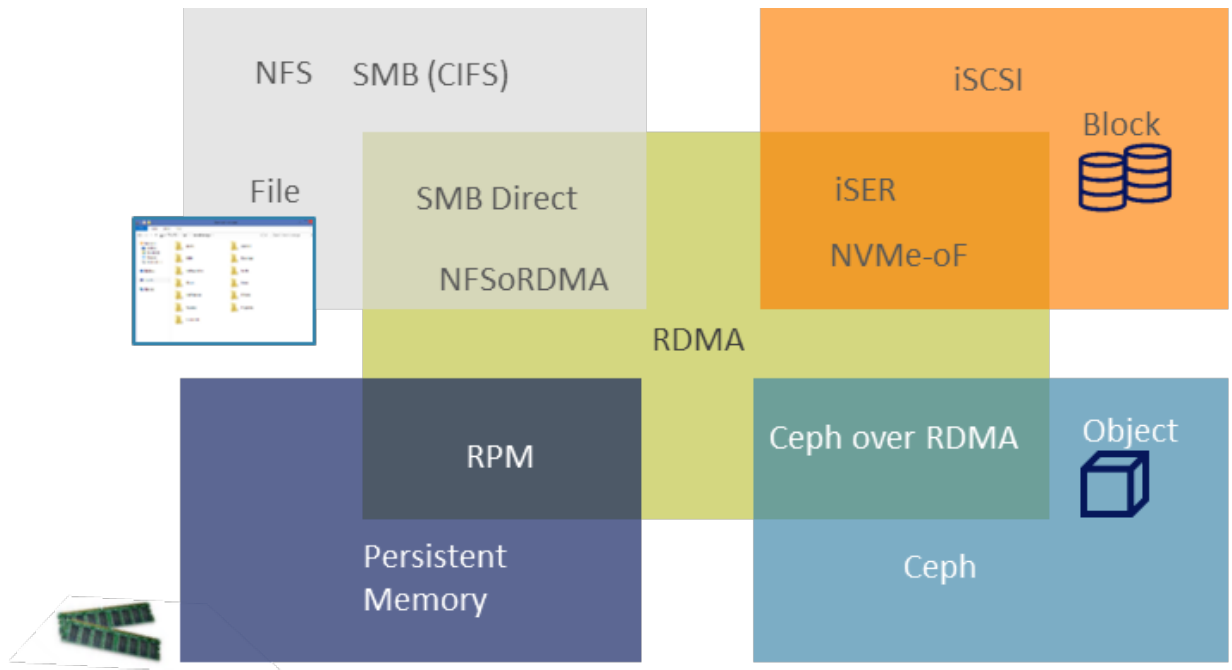
## With RDMA

- 2x Better Bandwidth
- Half the Latency
- 33% Lower CPU

See MS demo: <https://www.youtube.com/watch?v=u8ZYhUjSUoI>

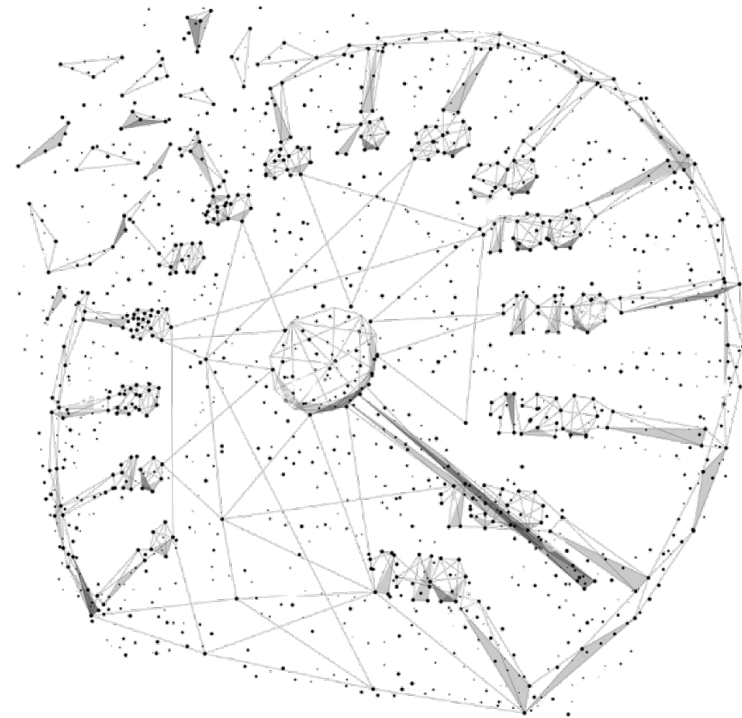
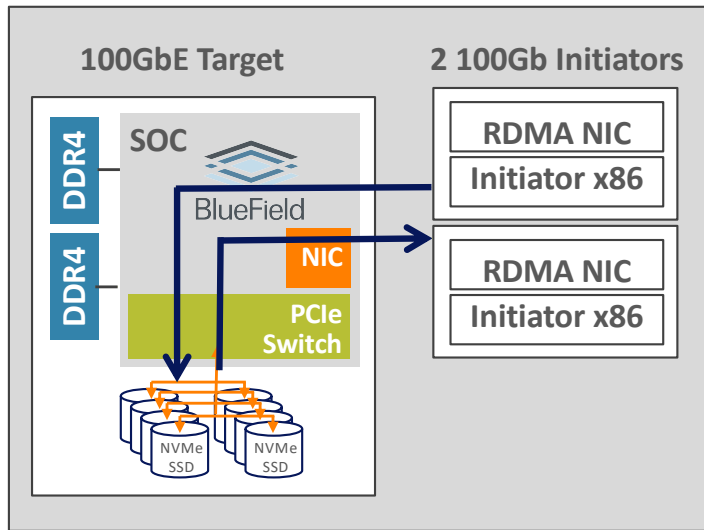


# Industry Wide RDMA Adoption



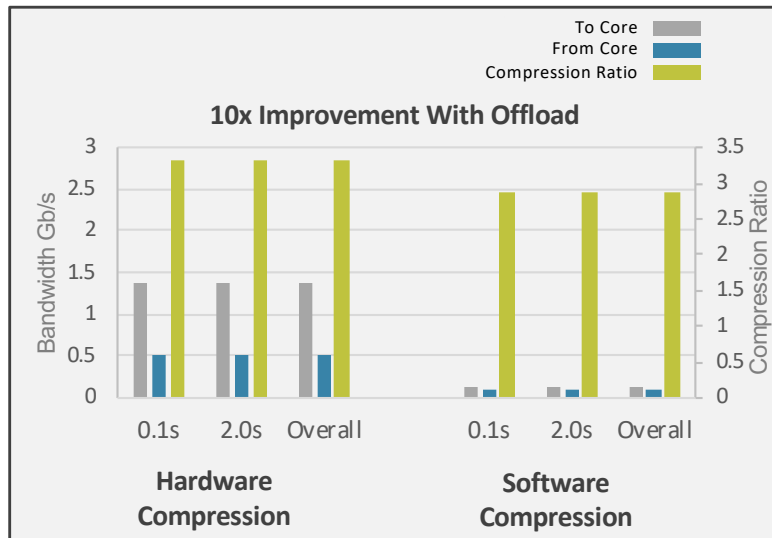
# NVMe-oF Performance with RDMA

- More Bandwidth     ✓ 5M IOPs, 4K block side
- Less Latency        ✓ ~3usec latency
- Less CPU             ✓ 0.01% CPU utilization

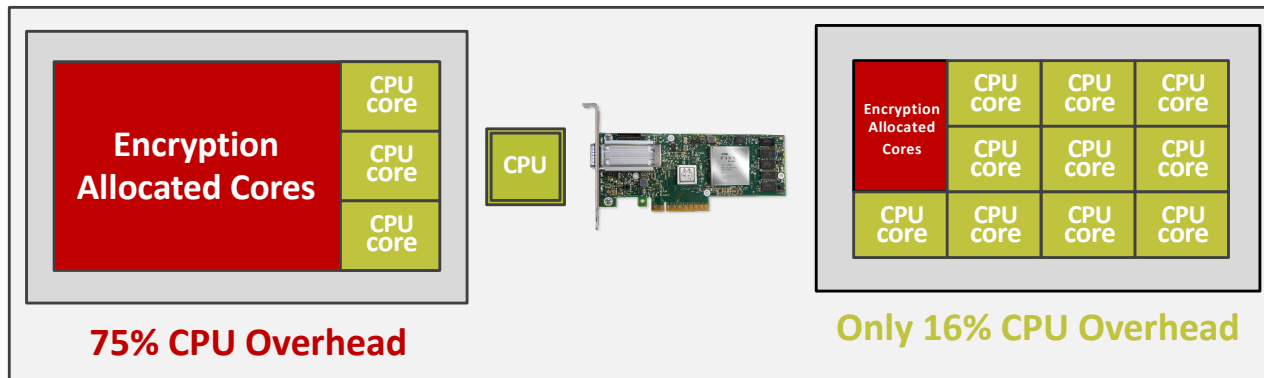




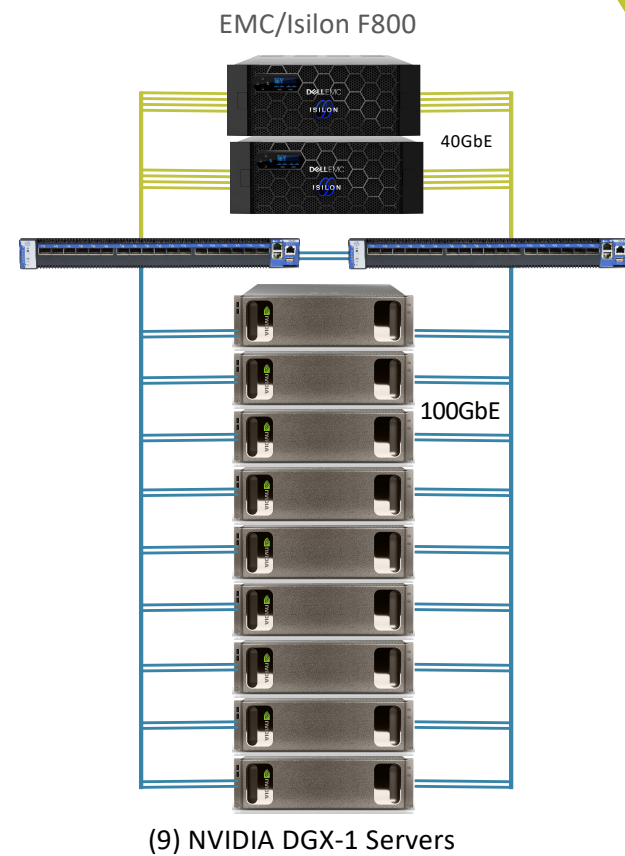
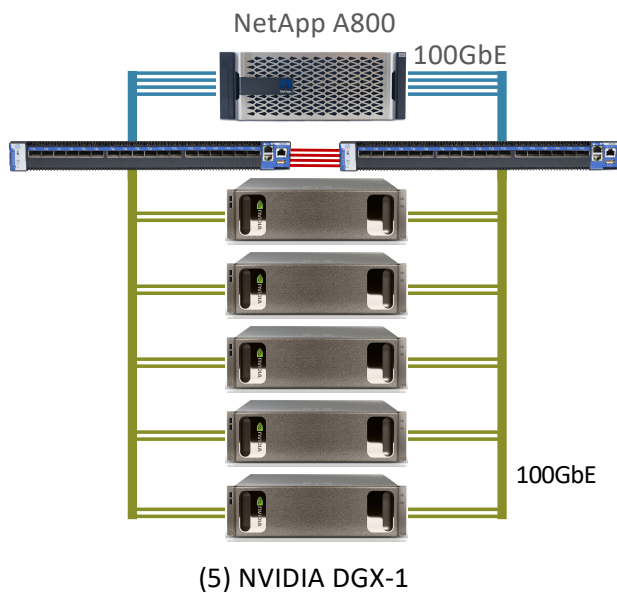
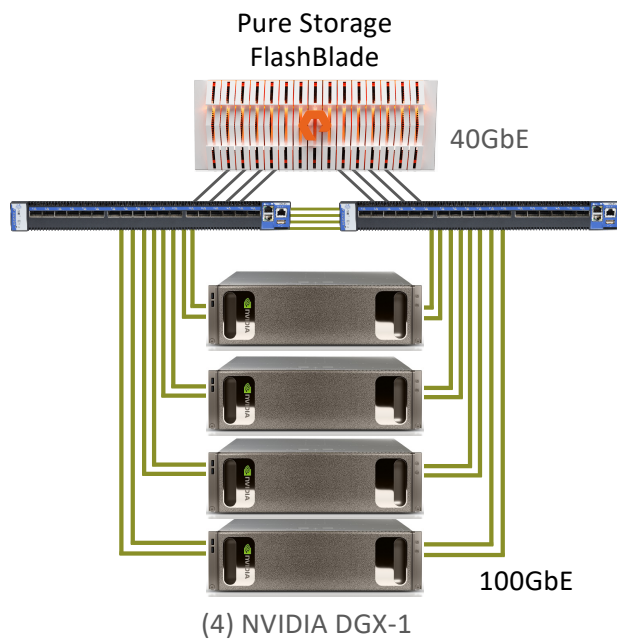
# Offloads



- Compression
- Security
- Others...



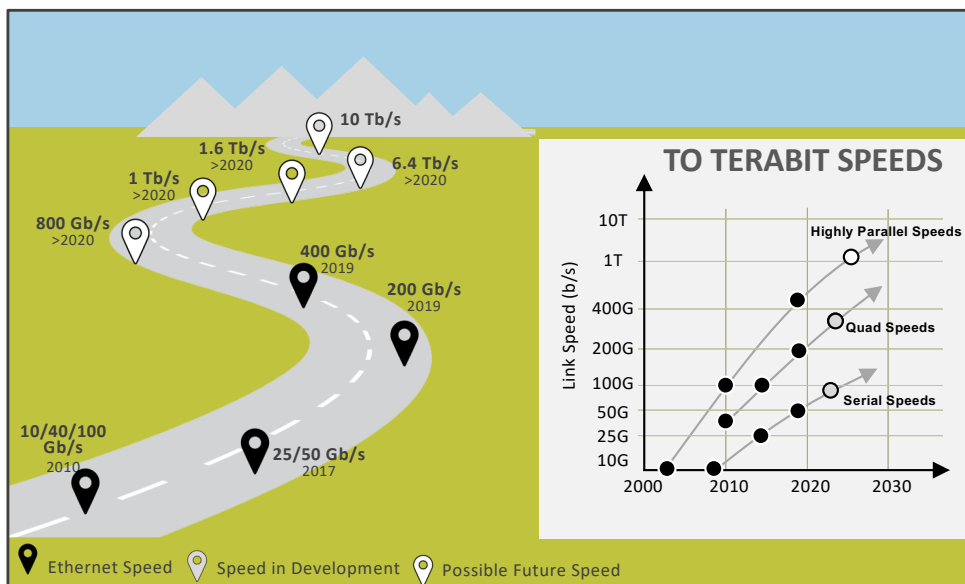
# Solutions



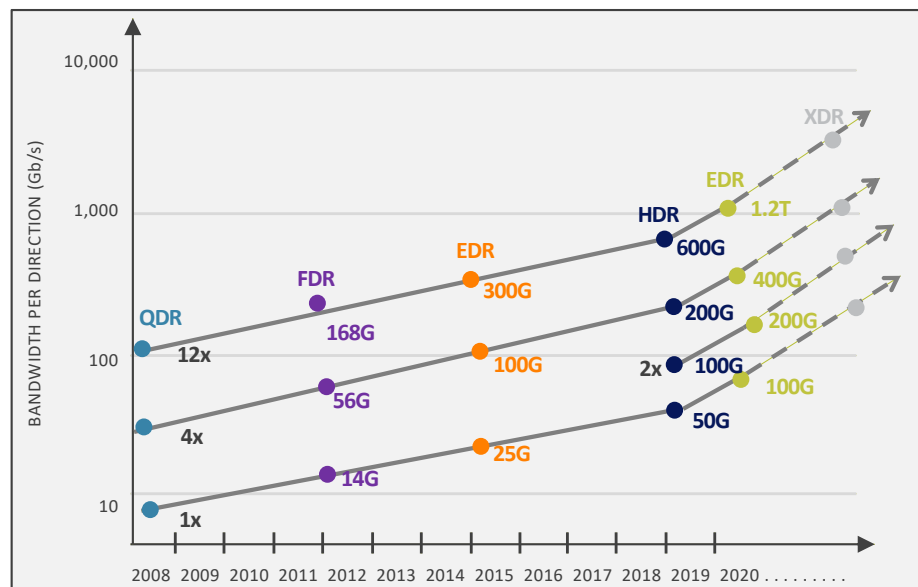
# Future Network Performance



## Ethernet

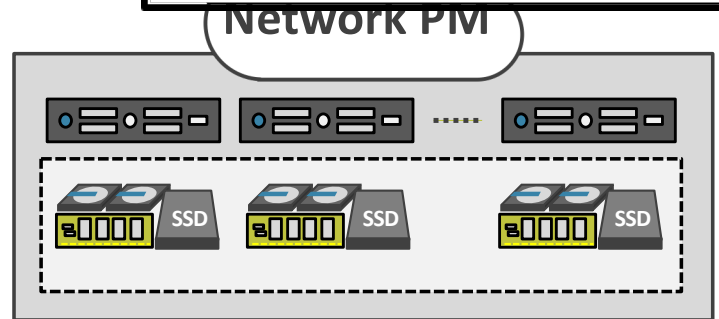
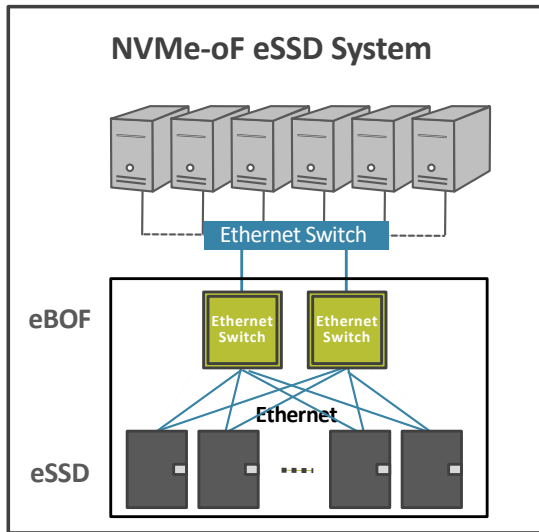
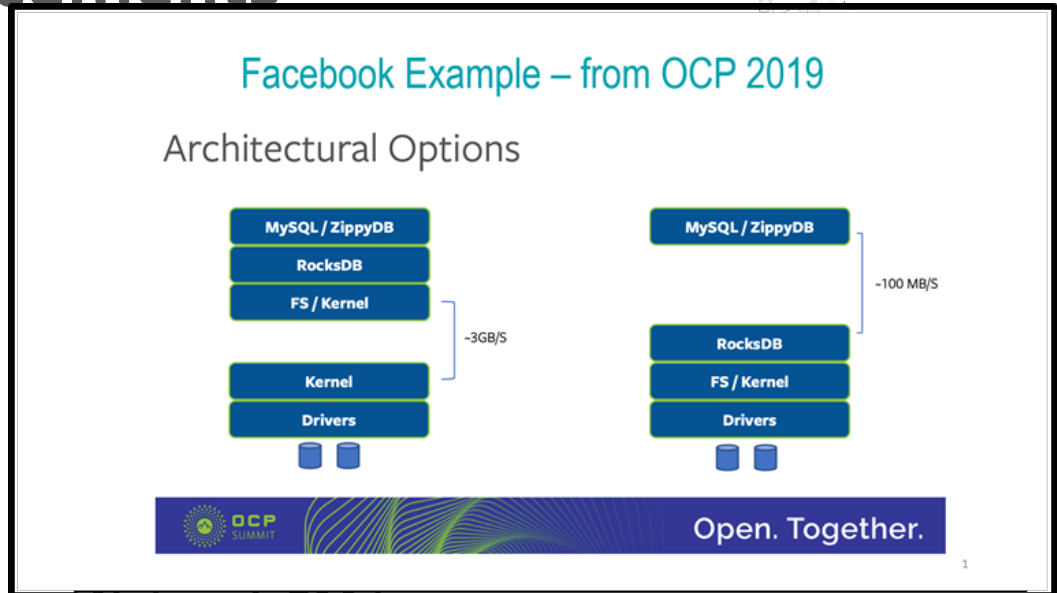


## InfiniBand



# Future Storage Advancements

- Ethernet SSDs
- Remote Persistent Memory (RPM)
- Computational Storage





Flash Memory Summit

**Thanks You!**



**Chris Lamb**

**Michael Kagan**





Flash Memory Summit

# Thanks You!



**Chris Lamb**  
**VP Compute Software at NVIDIA**  
**Michael Kagan**  
**CTO at Mellanox**

